STANFORD DIGITAL REPOSITORY -- FORMAT SCORING MATRIX

The Format Scoring Matrix (below) provides a scale for measuring the long-term sustainability of a digital file format or group of like formats. It is based in large part on the work of the Library of Congress Office of Strategic Initiatives. Formats are evaluated against the five factors listed in the left columns (definitions for the factors appear on page 2). The format scores positively for a given factor if that factor is considered not to impede future digital preservation efforts; a negative score is assigned when the factor is expected to hinder preservation. The accumulation of negative scores results in an increasingly lower status for a format. Therefore, a higher score indicates that a format's prospects for sustainability are poorer than one with a low score. If no match on the matrix is made, or if the MIME type is determined to be application/octet-stream, a file undergoign assessment receives a score of 5. The resulting Format Score is matched to a corresponding Preservation Quality Value (see table to right).

Preservation Quality Values:

Neg. Count	Status
0	high
1	
2	medium
3	
4	low
5	

	Text		**Image/Graphic**							**Page-viewer**
	Plain (ASCII, UTF-8)	Marked-up (html, css, xml, sgml)	JPEG	GIF	TIFF	JPEG2000 (lossless comp)	BMP	PNG	Photoshop	PDF
Adoption	+	+	+	+	+	-	+	-	+	+
Disclosure	+	+	+	+	+	+	+	+	-	+
Transparency	+	+	+	-	+	+	+	+	-	-
Self-Documentation	+	+	-	-	+	+	+	+	-	+
External Dependencies	+	-	+	+	+	+	+	+	-	-
Negative Count	0	1	1	2	0	1	0	1	4	2

	Office Docs	**Audio**			**Video**				**Animation**
	Word, Excel, PowerPoint	WAVE (LPCM)	AIFF	MP3	MPEG	Real	QT	Windows Media	Flash
Adoption	+	+	+	+	+	+	+	+	+
Disclosure	-	+	+	+	+	-	-	-	-
Transparency	-	+	+	-	-	-	-	-	-
Self-Documentation	-	+	+	+	+	-	-	-	-
External Dependencies	-	+	+	+	+	-	-	-	-
Negative Count	4	0	0	1	1	4	4	4	4

The matrix has been formatted to fit on a single printed page.

12/16/2004

The following definitions for sustainability factors -- "factors that influence feasibility and cost of preserving content in the face of future change" -- have been excerpted from "Digital Formats: Factors for Sustainability, Functionality, and Quality", a presentation by Caroline Arms and Carl Fleischhauer, Office of Strategic Initiatives, Library of Congress, at the Digital Library Federation Forum in Fall 2003 (available online at: http://memory.loc.gov/ammem/techdocs/digform/ and http://www.diglib.org/forums/fall2003/fallforum03.htm#p1)

Disclosure	Degree to which complete specifications and tools for validating technical integrity exist and are accessible to those creating and sustaining digital content. A spectrum of disclosure levels can be observed for digital formats. What is most significant is not approval by a recognized standards body, but the existence of complete documentation. Preservation of content in a given digital format over the long term is not feasible without an understanding of how the information is represented (encoded) as bits and bytes in digital files.
Adoption	Degree to which the format is already used by the primary creators, disseminators, or users of information resources. This includes use as a master format, for delivery to end users, and as a means of interchange between systems. If a format is widely adopted, it is less likely to become obsolete rapidly, and tools for migration and emulation are more likely to emerge from industry without specific investment by archival institutions.
Transparency	Degree to which the digital representation is open to direct analysis with basic tools, such as human readability using a text-only editor. Digital formats in which the underlying information is represented simply and directly will be easier to migrate to new formats and more susceptible to digital archaeology; easier development of rendering software for new technical environments.
Self-documentation	Self-documenting digital objects contain basic descriptive, technical, and other administrative metadata. Self-documenting digital objects are likely to be easier to sustain over the long term and to transfer reliably from one archival system to another, including a successor system. LC wants to take advantage of the trend towards embedded metadata for business reasons. Some metadata will be extracted to support discovery and collection management.
External Dependencies	Degree to which a particular format depends on particular hardware, operating system, or software for rendering or use and the predicted complexity of dealing with those dependencies in future technical environments. Some interactive digital content is designed for use with specific hardware, such as a joystick. Scientific datasets built from sensor data may require specialized software for analysis and visualization. External dependencies will make content more difficult and costly to sustain than static content. The specialized software required by some scientific datasets may itself be very difficult to sustain.